FEBS 18381 FEBS Letters 406 (1997) 69–74

Distribution of sequence-dependent curvature in genomic DNA sequences

Andrei Gabrielian[1], Kristian Vlahovicek, Sándor Pongor*

International Centre for Genetic Engineering and Biotechnology (ICGEB), Area Science Park, Padriciano 99, 34012 Trieste, Italy

Abstract The distribution of inherent, sequence-dependent curvature was calculated for a number of prokaryotic (*M. genitalium, H. influenzae, M. jannaschii*), viral (adenovirus 2, equine herpes virus 1), phage (M13, lambda), eukaryotic (*S. cerevisiae*) and mitochondrial genomes as well as *E. coli* and human genomic fragments. The genomic averages are in the range of 6–8°/helical turn and only about 20% of DNA is curved less than 3°/helical turn. The prokaryotes and phages appear to have a consistently higher frequency of curved DNA in their genomes than the other genomes tested. Long, highly curved segments, similar to artificially designed curved DNA, are apparently absent from the genomes. Short, curved segments, differing in G+C content may provide environmentally modulated conformational signals for gene regulation. A WWW-server was constructed for the prediction of curved sites from DNA sequences ⟨http://icgeb.trieste.it/dna/curve_it.html/⟩..

© 1997 Federation of European Biochemical Societies.

Key words: DNA structure; DNA curvature; Genome analysis

1. Introduction

The fact that DNA may contain local, sequence-dependent conformations has profoundly influenced the thinking of biologists in recent years. DNA is no longer considered to be an idealized, featureless double helix but rather a series of local conformations that may add up to macroscopic curvature even without external factors such as DNA-binding proteins [1]. Inherent bending of DNA is usually associated with periodically repeating sequence motifs such as A-tracts [2] or GC-type elements [3]. In this respect, it is customary to speak about A-type and non-A-type DNA curvature that differ in their sensitivity to metal ions [3] and in the direction of bending [4]. Given the differential response of curved motifs to environmental factors, it is plausible to speculate that these motifs may act as 'environmental sensors' whose conformational transitions act as regulatory signals. The aim of the present work is to obtain preliminary information on how inherently curved DNA segments are distributed in genomic sequences and to see if the currently available complete genomes differ in this respect.

Prediction of DNA curvature is based on the geometry of the individual dinucleotide steps (roll, twist and tilt angles). There are several methods for the calculation of the DNA trajectory that may give different results, especially for longer DNA segments [5–9]. However, for shorter segments, like those involved in local bending, essentially all of these are suitable, so we chose a simple and visually meaningful index, the angle of basepairs one helical turn apart (schematically shown on the model of the d(CGCGAATTCGCG) double helix [10], Fig. 2). This value is expected to be zero for ideally straight B-DNA and positive for curved segments.

Here we show, through the analysis of complete genomes, that prokaryotic genomes (*H. influenzae*, *M. jannaschii*, *M. genitalium*) appear to contain more curved segments than the human, yeast and viral genomic DNA tested. Long, uniformly curved segments, such as can be easily produced with artificially designed repetitive sequence motifs, seem to be absent in genomic DNA. However, shorter, curved segments differing in G+C contents are ubiquitous.

2. Methods

There are a number of published geometric parameter sets determined on the basis of X-ray data [11], theoretical geometry calculations [7], gel mobility analysis [12] as well as NMR measurements [13] of which we used the latter. The magnitude of DNA curvature was calculated with the BEND algorithm of Goodsell and Dickerson [5,14]. In this algorithm, the normal vectors of successive basepairs are added up vectorially and the angle between two averaged normal vectors 31 basepairs (~3 helical turns) apart is used as the indicator of curvature. We presented the values of the curvature as the deflection angle per 10.5 residue helical turn (1°/bp=10.5°/helical turn). The genomic sequences were taken from the sequence databases and from the literature, as indicated. The distribution of curvature is presented as histograms with a bin-width of 1.5°/helical turn, i.e. the percentage of segments with a curvature between 0 and 1.5°/helical turn is plotted at 0°/helical turn, etc. The average distributions (Fig. 2) are arithmetic averages of the groups presented in Fig. 1A-D, respectively.

3. Results and discussion

Even though curvature values determined by gel mobility analysis strongly depend on the experimental conditions [15] and on the length of the tested molecule [16], there are a number of sequence motifs whose curvature characteristics were established by several research groups, and we used these to test the chosen method of calculation. The results in Table 1 show that there is a good over-all agreement between the curvature index (angular deflection per helical turn) with the experimental results. The calculated values for curved DNA is above 9°/helical turn, the straight sequence motives give values below 4–5°/helical turn. There are exceptions to this rule (i.e. wrong predictions); e.g. the $(tctctaaaaaatatataaaaa)_n$ motif, which is highly curved according to electrophoretic mobility and circulatization kinetics experiments [17], gives a conspicuously low value. Therefore this calculation — like all predictions — has to be considered approximate, so absolute threshold values may not be defined. Nevertheless, the correlation with experimental curvature values in these and in previously reported examples [14] is satisfactory. Roughly

*Corresponding author.
E-mail: pongor@icgeb.trieste.it

[1] *Permanent address:* Institute of Molecular Biology, Russian Academy of Sciences, Vavilov St. 32, 117984 Moscow, Russia.

0014-5793/97/$17.00 © 1997 Federation of European Biochemical Societies. All rights reserved.
PII S0014-5793(97)00236-6

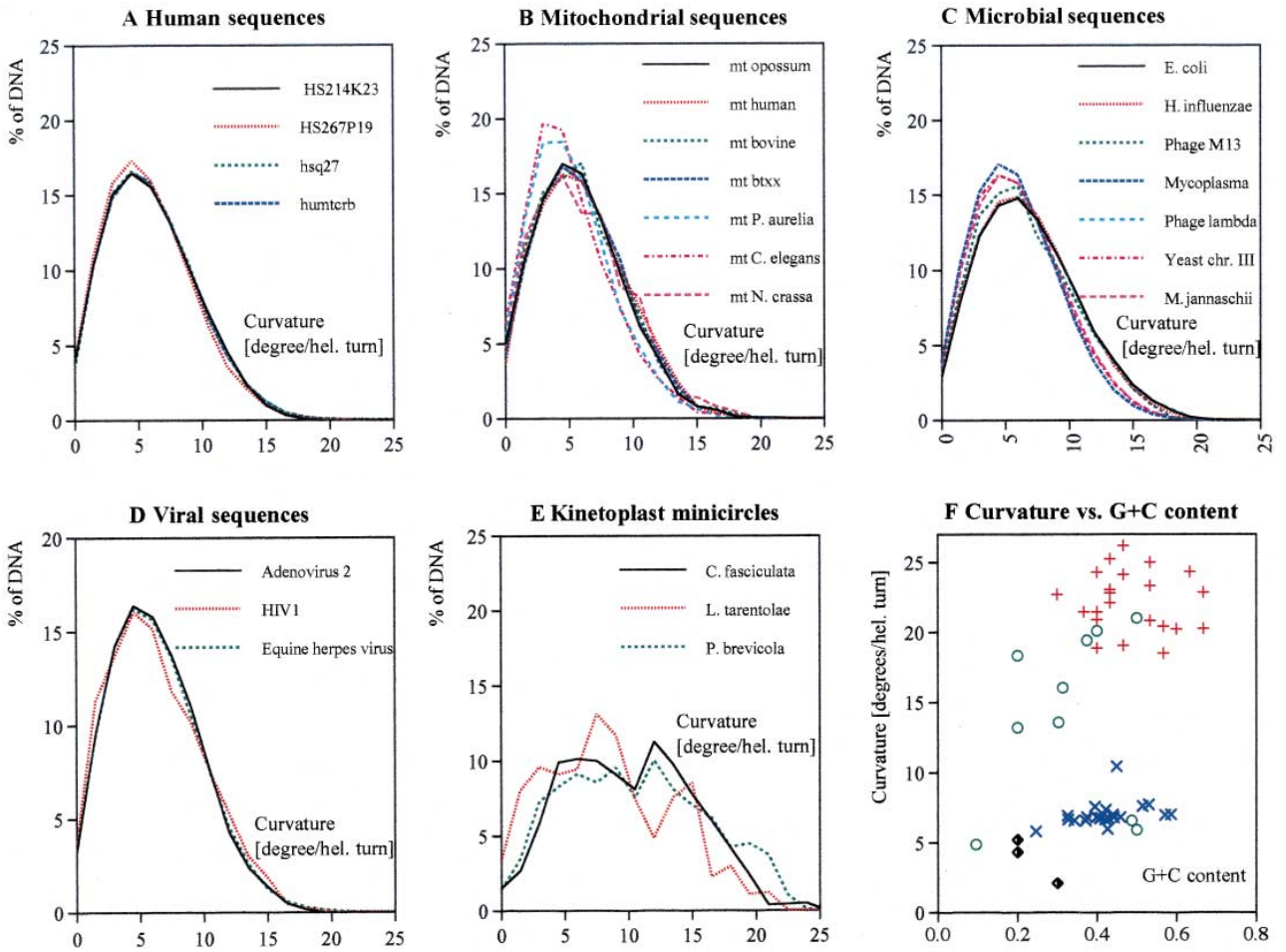

Fig. 1. Distribution of curvature in genomic DNA sequences. The curvature is characterized by the angle of deflection per 10.5 bp helical turn (X-axis). The references of the DNA sequences are under Table 2. A: Human sequences; B: mitochondrial sequences; C: microbial sequences; D: viral sequences; E: kinetoplast minicircle DNA sequences (GenBank locus name, length given in parentheses) *C. fasciculata*, 1371 bp; KILTB4MC_S2. *L. tarentolae*, 887 bp; LDKMPL13_S7, *P. brevicola*, 1477 bp; TBU03908_S2; F: correlation of curvature with G+C content in genomic averages (blue x), genomic maxima (red +) curved synthetic oligonucleotides (green o) and straight synthetic oligonucleotides (black diamonds). The calculated correlation coefficients are below 0.15 in each group.

speaking, we considered values above 15°/helical turn as curved.

The average curvature in genomic sequences (Table 2) is between 6.0 and 7.7°/helical turn, even though less than 20%

Table 1
DNA curvature[a] calculated for straight and curved DNA sequence motifs

No.	Origin (reference)	Sequence	G+C content	Curvature[b] (°/helical turn)
Curved DNA				
1	Synthetic[21]	$(aaaattttgc)_n$	0.200	18.333
2	Synthetic[21]	$(aaaattttcg)_n$	0.200	13.193
3	Synthetic[22]	$(tctcaaaaaacgcgaaaaaaccggaaaaaagc)_n$	0.375	19.430
4	Synthetic[23]	$(ccgaaaaagg)_n$	0.500	20.999
5	Synthetic[24]	$(tctctaaaaaatatataaaaa)_n$	0.095	4.880
6	Synthetic[23]	(ggcaaaaaac)n	0.400	20.093
7	*L.tarentolae* kinetoplast[25]	ccaaaaatgtcaaaaaataggcaaaaaatgcc	0.313	16.027
8	Synthetic[25]	aaaaactctctaaaaactctccctagaggggccctagagggc	0.500	5.893
9	Synthetic[25]	aaaaactctctaaaaactctaagaggggccctagagggccc	0.488	6.579
10	*C.risortia* bent satellite DNA[26]	agaattgggacaaaaattggaaatttttaaggg	0.303	13.579
Straight DNA				
11	Synthetic[27]	$(atctaatctaacacaacaca)_n$	0.300	2.091
14	OR3 operator region[28,29]	actacgttaaatctatcaccgcaagggataaa	0.375	5.193
15	OR3 region, mutated[29,30]	actacgttaaatctatcaccacaagggataaa	0.344	4.302
16	poly-A[30]	$(a)_n$	0.000	0.008

[a]The angular deflection [degree per helical turn] was determined with the BEND algorithm [1] using a window size w = 31 (~3 helical turns) and the values were corrected to a helical repeat length of 10.5 nucleotides.
[b]By definition, homopolymers should give zero curvature. The non-zero value indicates the numeric precision of the calculation.

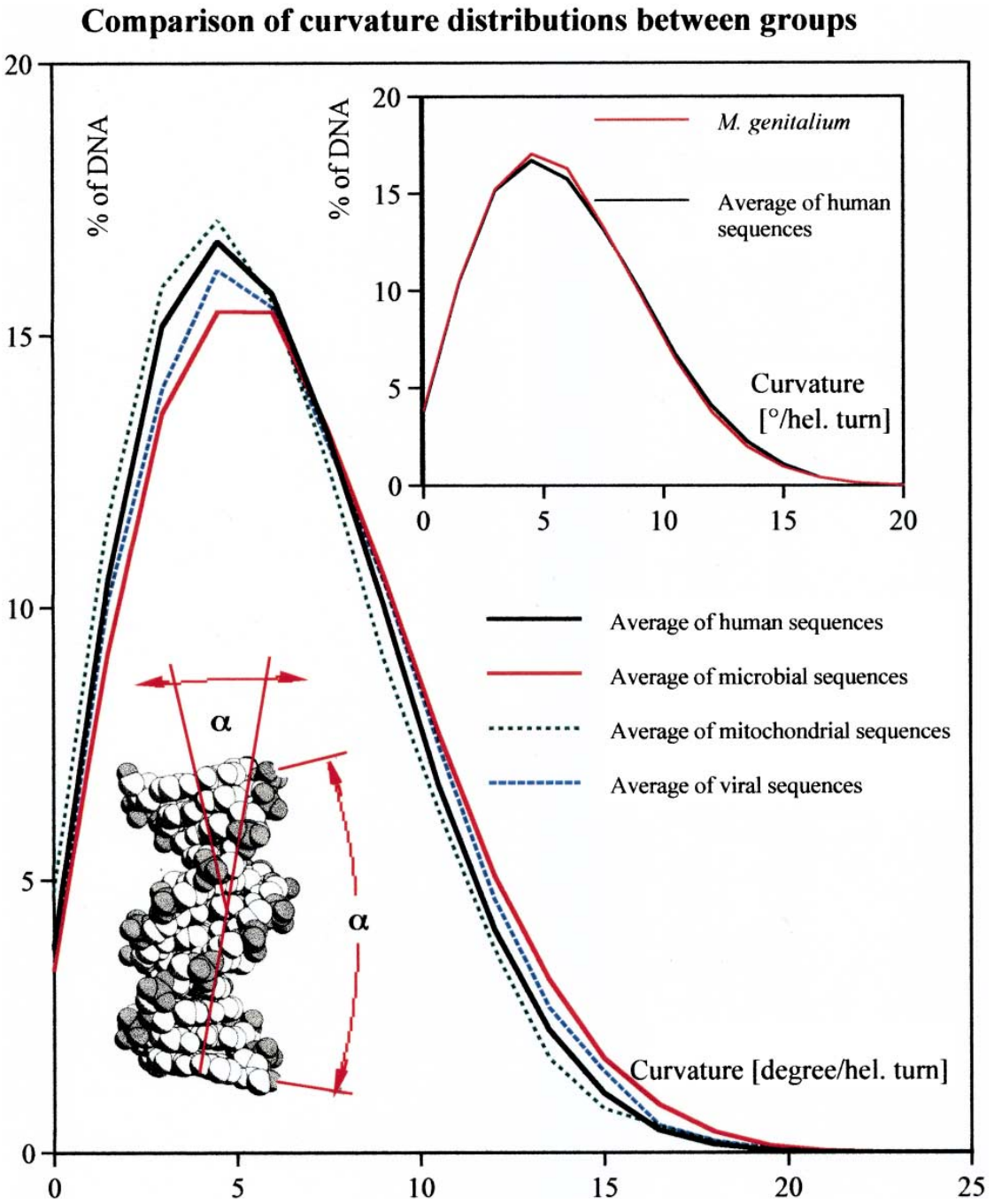

Fig. 2. Distribution of curvature in various genomic groups. The average distributions were calculated in the groups presented in Fig. 1. The upper inset shows the comparison of the *M. genitalium* genome with the human group. The lower inset is a schematic representation of the curvature angle using the model of the d(CGCGAATTCGCG) double helix [10]. The references of the DNA sequences are under Table 2.

of DNA is below the limit of 3°/helical turn. This in accordance with the intuitive expectation that average DNA is reasonably straight if not exposed to external factors. Short sequence segments may have very different average values, as shown by the example of the *L. tarentolae* kinetoplast minicircle (Table 2). The maximum values found in the genomes (Table 2, column 5) are quite similar to those found with artificially designed sequence motifs (Table 1, column 5). However, it is conspicuous, that the longest stretches of continuous curvature (Table 2, column 8) do not reach the length of 100–200 bp, i.e. the oligonucleotide length necessary for detecting curvature by gel mobility analysis [18]. A similar conclusion was reached previously based on DNA bendability statistics [19].

At first sight, the distribution of curvature appears quite similar in all the genomes tested (Fig. 1). It follows a typical, even though not symmetrical random distribution. The distribution is smooth, apparently there are no preferred values of curvature. The human genomic segments and the viral genomes (Fig. 1A,B, respectively) show quite similar distributions among themselves, while mitochondrial and prokaryotic genomes (Fig. 1C,D, respectively) are apparently more varia-

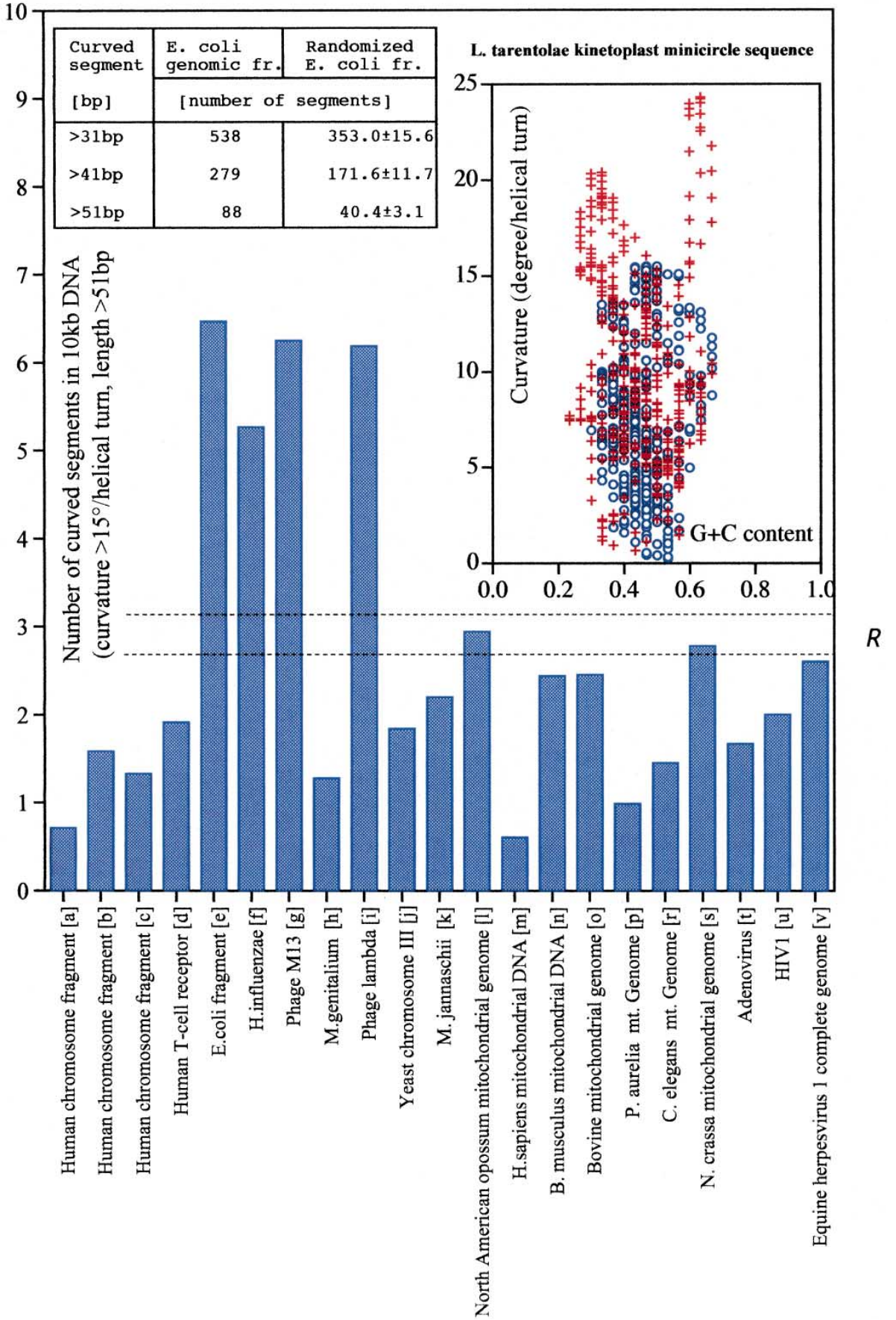
Curved segment [bp]
E. coli genomic fr. [number of segments]
Randomized E. coli fr.
>31bp 538 353.0±15.6
>41bp 279 171.6±11.7
>51bp 88 40.4±3.1
L. tarentolae kinetoplast minicircle sequence
Curvature (degree/helical turn)
G+C content
Number of curved segments in 10kb DNA (curvature >15°/helical turn, length >51bp
R
Human chromosome fragment [a]
Human chromosome fragment [b]
Human chromosome fragment [c]
Human T-cell receptor [d]
E.coli fragment [e]
H.influenzae [f]
Phage M13 [g]
M.genitalium [h]
Phage lambda [i]
Yeast chromosome III [j]
M. jannaschii [k]
North American opossum mitochondrial genome [l]
H.sapiens mitochondrial DNA [m]
B. musculus mitochondrial DNA [n]
Bovine mitochondrial genome [o]
P. aurelia mt. Genome [p]
C. elegans mt. Genome [r]
N. crassa mitochondrial genome [s]
Adenovirus [t]
HIV1 [u]
Equine herpesvirus 1 complete genome [v]

Fig. 3. Frequency of contiguous curved segments found in various genomes, given as number of curved segments with a curvature above 15°/helical turn/10 kb DNA. *R*, the range of curved segment frequency found in random shuffled genomic sequences. The table (inset) shows that *E. coli* DNA (136.1 kb, Genbank ECUW87) contains significantly more curved segments than its random shuffled counterpart (mean and variance of five random shufflings). The right inset shows that the characteristic curvature vs. G+C pattern of the *L. tarentolae* minicircle DNA (red, 887 bp; GenBank LDKMPL13_S7) disappears on random shuffling (blue).
←

ble. While all of these distributions are roughly similar, kinetoplast DNA shows a characteristically different distribution, due to the fact that curved elements make up a high proportion of the relatively short sequences (Fig. 1E).

Are there specific differences between genomes? The comparison of average distributions in Fig. 2 shows that microbial genomes, on the average, contain more curved segments than the other genome groups (human, viral and mitochondrial). The microbial group is not homogenous (see also Fig. 1C). For example, the genome of the human parasite *M. genitalium*, an organism with a very reduced genome adapted to the human organism's environment, is in fact closer to the human sequences than to the rest of the microbial group (Fig. 2, inset), and so is the yeast chromosome III sequence. Mitochondrial genomes, on the other hand, are much closer to the human and the viral groups, than to the microbial genomes. The number of contiguous curved segments (Fig. 3) shows roughly similar tendencies: only prokaryotic genomes, with the exception of *M. genitalium*, apparently contain more such segments than viral and human genomic DNA. The genomes also differ in the number of long, curved segments (> 15°/helical turn). The frequency of such segments is above 5/10 kb DNA in the bacterial and phage sequences tested (*E. coli*, 6.5, *H. influenzae*, 5.3, Phage lambda, 6.2; Phage M13, 6.3), except *M. genitalium*, in which this value is 1.3, close to the average of the human sequences The same value is typically below 3.0 in the human, yeast and viral sequences.

A plausible question is whether the predicted curvature values depend on the bias of the DNA composition such as an extreme G+C content. This can be expected since the example on which the calculation apparently failed, is a motif of low G+C content (9.5%, Table 1, column 5), and the prokaryotic genome of lowest apparent curvature, *M. genitalium*, has the lowest G+C content. In the sequences analyzed here, however, one sees no apparent correlation between curvature and G+C content (Fig. 1F). Also, the genomic maxima of curvature are associated with a relatively wide range of G+C content. This is not entirely surprising, on the other hand, since curvature, as calculated here, depends on the serial order of the dinucleotides, rather than on the dinucleotide composition of the sequence. Another plausible question to ask is whether patterns of curvature may occur in random sequences. This is apparently not the case; the distribution characteristic of curved DNA 'collapses' upon random shuffling of the sequence (Fig. 3, insets). It is interesting to note, that random shuffling of different sequences of varying G+C content yields apparently identical curvature distributions (not shown), and

Table 2
Distribution of DNA curvature[a] (°/helical turn] in genomic DNA[b–w]

Genomic DNA	Size	Average curvature (SD; °/hel.turn)	Average G+C content (SD)	Max. curvature (°/hel.turn; content)	% below 3°/hel.turn	% above 15°/hel.turn	Longest segment above 15°/hel.turn (bp)
Human chromosome fragment[b]	127 kbp	6.751 ± 3.459	0.409	20.821 (0.533)	14.30	1.46	27
Human chromosome fragment[c]	113.7 kbp	6.560 ± 3.385	0.371	21.427 (0.400)	14.86	1.42	39
Human chromosome fragment[d]	37.7 kbp	6.812 ± 3.498	0.375	24.117 (0.467)	13.69	1.93	34
Human T-cell receptor[e]	350 kbp	6.798 ± 3.517	0.442	23.319 (0.533)	14.15	1.97	41
E. coli fragment[f]	136.1 kbp	7.699 ± 3.943	0.531	25.231 (0.433)	10.81	4.73	56
H. influenzae[g]	1830 kbp	7.561 ± 3.828	0.394	26.188 (0.467)	11.30	3.93	56
Phage M13[h]	6.4 bp	7.330 ± 3.798	0.422	20.264 (0.667)	12.05	3.36	33
M. genitalium[i]	580 kbp	6.652 ± 3.399	0.327	22.703 (0.300)	14.34	1.53	49
Phage lambda[j]	48.5 kbp	7.594 ± 3.883	0.515	22.105 (0.433)	11.28	4.31	37
Yeast chromosome III[k]	315.3 kbp	6.799 ± 3.529	0.398	23.034 (0.433)	14.27	1.98	49
M. jannaschii[l]	1665 kbp	6.913 ± 3.537	0.326	24.285 (0.400)	13.51	2.10	59
North Am. opossum mt. genome[m]	17 kbp	6.566 ± 3.460	0.343	22.807 (0.433)	15.54	1.65	26
H. sapiens mitochondrial genome[n]	16.5 kbp	6.808 ± 3.463	0.459	20.220 (0.600)	14.36	1.32	22
B. musculus mitochondrial DNA[o]	16.4 kbp	6.639 ± 3.477	0.420	20.879 (0.400)	15.30	1.76	40
Bovine mitochondrial genome[p]	16.3 kbp	6.725 ± 3.448	0.407	18.858 (0.400)	14.47	1.52	33
P. aurelia mt. Genome[q]	40.5 kbp	5.944 ± 3.209	0.426	21.431 (0.367)	18.60	0.97	25
C. elegans mt. Genome[r]	13.8 kbp	5.797 ± 3.223	0.246	18.486 (0.567)	19.61	0.81	31
N. crassa mitochondrial genome[s]	3.6 kbp	6.709 ± 3.665	0.428	19.041 (0.467)	16.48	2.60	24
Adenovirus 2[t]	36 kbp	6.957 ± 3.479	0.571	22.843 (0.667)	12.75	2.02	29
HIV-1[u]	10 kbp	6.980 ± 3.681	0.440	20.405 (0.567)	14.89	2.54	30
Equine herpes virus 1[v]	150 kbp	6.985 ± 3.592	0.586	24.985 (0.533)	13.09	2.28	44
Leishmania kinetoplast[w]	414 bp	10.429 ± 5.125	0.449	24.323 (0.633)	4.35	18.12	29

[a]The angular deflection (°/helical turn) was determined with the BEND algorithm 1 using a window size w = 31 (~3 helical turns) and the values were corrected to a helical repeat length of 10.5 nucleotides.
[b]EMBL HS214K23; [c]EMBL HS267P19; [d]EMBL hsq27; [e]Genbank HUMTCR; [f]Genbank ECUW87; [g]TIG HIDB; [h]Genbank INM13X; [i]TIGR MGDB; [j]Genbank LAMCG; [k]Genbank SCCHRII; [l]TIGR MJDB; [m]GenBank dvntgme; [n]EMBL hsmitg; [o]Genbank MIBMCG; [p]Genbank MIBTXX; [q]MIPAGEN; [r]Genbank MTCE; [s]Genbank NEUMTPLCG; [t]Genbank ADRCG; [u]K03455; [v]Genbank HSECOMGEN; [w]Genbank KILTB4MC.

the number of curved segments found in them is quite uniform (R in Fig. 3).

Since the curvature distributions presented here are derived from a very large number of data (approximately as many as the length of the genomes), we think that these tendencies may not be artefactual. We note, that we have repeated the above calculations using the dinucleotide helical parameters of Gorin et al. [11] and of Bolshoy et al. [12], based on X-ray crystallography and gel-electrophoresis data, respectively. Even though the absolute values of the predicted curvature seems to depend on the model, the same general tendencies were found and the distribution of the data was remarkably similar in all cases. It has to be noted that the scarcity of long, uniformly curved segments in natural DNA does not preclude the possibility that curved segments serve as biological signals. On the contrary, the continuous curved segments (Fig. 3) are plausible candidates for functional analysis. It has to be noted, in addition, that even moderate curvature is believed to alleviate DNA-protein docking [20]. Gene regulation of prokaryotes relies predominantly on such contacts while eukaryotic systems frequently use additional protein–protein interactions. More detailed studies will be necessary to decide whether this or similar reasons are behind the elevated level of curvature in prokaryotic DNA. The facts that the genomic maxima have quite different G+C content, and that AA-type and GC-type elements are well known to react very differently to environmental conditions, such as ions [3], suggest that curved elements in genomic DNA may take part in differential structural rearrangements in response to changes in the environment. One can speculate that curved elements which differ in their flexibility [19] and in their responses to ions [3] may act as finely tuned sensors providing modulatory signals for gene-regulation processes.

Summarizing we can conclude that long, curved DNA segments that are necessary to detect curvature in vitro by gel mobility assay, or similar to those present in kinetoplast minicircles, are hardly found in genomic sequences. Even shorter, e.g. 50 bp long curved segments are rare, typically less than three such segment occur in 10 kb. However, less than 20% of the genomes have a predicted curvature below 3°/helical turn. In other words, moderate curvature in shorter segments is quite frequent, and, within the genomes analyzed in this work, prokaryotic DNA seem to contain a consistently higher number of these than do the higher organisms and viruses. Finally we mention that a WWW server allowing for the prediction of curved sites through curvature vs. G+C content plots (Fig. 3, inset) or curvature vs. sequence plots (not shown) is now publicly accessible (http://icgeb.trieste.it/curve_it.html/).

Acknowledgements: The advice of Profs. A. Falaschi and F.E. Baralle, and the help of Ms. S. Kerbavcic with the manuscript are gratefully acknowledged.

References

[1] Travers AA, Klug A. In: Cozzarelli NR, Wang JC, editors. DNA Topology and Its Biological Effects. New York: Cold Spring Harbor Laboratory, 1990:57–106.
[2] D.M. Crothers, T.E. Haran, J.G. Nadeau, J Biol Chem 265 (1990) 7093–7096.
[3] I. Brukner, S. Susic, M. Dlakic, A. Savic, S. Pongor, J Mol Biol 236 (1994) 26–32.
[4] I. Brukner, M. Dlakic, A. Savic, S. Susic, S. Pongor, D. Suck, Nucl Acids Res 21 (1993) 1025–1029.
[5] D.S. Goodsell, R.E. Dickerson, Nucl Acids Res 22 (1994) 5497–5503.
[6] E.S. Shpigelman, E.N. Trifonov, A. Bolshoy, Comp Appl Biosci 9 (1993) 435–440.
[7] P. De Santis, A. Palleschi, M. Savino, A. Scipioni, Biochemistry 29 (1990) 9269–9273.
[8] R. Lavery, H. Sklenar, J Biomol Struct Dynam 6 (1989) 655–667.
[9] M.A. El Hassan, C.R. Calladine, J Mol Biol 251 (1995) 648–664.
[10] R. Wing, H. Drew, T. Takano, C. Broka, K. Tanaka, K. Itakura, R.E. Dickerson, Nature 287 (1980) 755–758.
[11] A.A. Gorin, V.B. Zhurkin, W.K. Olson, J Mol Biol 247 (1995) 34–48.
[12] A. Bolshoy, P. McNamara, R.E. Harrington, E.N. Trifonov, Proc Natl Acad Sci USA 88 (1991) 2312–2316.
[13] N. Ulyanov, T. James, Methods Enzymol 261 (1995) 90–115.
[14] A. Gabrielian, S. Pongor, FEBS Lett 393 (1996) 65–68.
[15] L.S. Shlyakhtenko, I. Liubchenko, B.K. Chernov, V.B. Zhurkin, Molekulyarnaya Biol. (russ) 24 (1990) 66–81.
[16] C.R. Calladine, C.M. Collis, H.R. Drew, M.R. Mott, J Mol Biol 221 (1991) 981–1005.
[17] L. Ulanovsky, M. Bodner, E. Trifonov, M. Choder, Proc Natl Acad Sci USA 83 (1986) 862–866.
[18] H.S. Koo, H.M. Wu, D. Crothers, Nature 320 (1986) 501–506.
[19] A. Gabrielian, A. Simoncsits, S. Pongor, FEBS Lett 393 (1996) 124–130.
[20] M. Suzuki, N. Yagi, J Mol Biol 255 (1996) 677–687.
[21] P.J. Hagerman, Nature 321 (1986) 449–450.
[22] C.R. Calladine, H.R. Drew, M.J. McCall, J Mol Biol 201 (1988) 127–137.
[23] H.S. Koo, H.M. Wu, D. Crothers, Nature 320 (1986) 501–506.
[24] L. Ulanovsky, M. Bodner, E. Trifonov, M. Choder, Proc Natl Acad Sci USA 83 (1986) 862–866.
[25] I. Brukner, M. Dlakic, A. Savic, S. Pongor, D. Suck, Nucl Acids Res 21 (1992) 1025–1029.
[26] Genbank CRBENSAT.
[27] J. Bednar, P. Furrer, V. Katritch, A. Stasiak, J. Dubochet, Mol Biol 254 (1995) 579–594.
[28] Y. Lyubchenko, L. Shlyakhtenko, B. Chernov, R.E. Harrington, Proc Natl Acad Sci USA 88 (1991) 5331–5334.
[29] Y. Lyubchenko, L.S. Shlyakhtenko, E. Appella, R.E. Harrington, Biochemistry 32 (1993) 4121–4127.
[30] M.A. Young, G. Ravishanker, D.L. Beveridge, H.M. Berman, Biophys J 68 (1995) 2454–2468.